Automatic content curation of news events

Hei-Chia Wang^{1,2} · Chun-Chieh Chen^{1,3} · Ting-Wei Li¹

Received: 11 April 2021 / Revised: 5 December 2021 / Accepted: 10 January 2022 /
Published online: 16 February 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

With the rapid development of the internet, a large amount of online news has brought readers a variety of information. Some important events last for some time as the event develops or the topic spreads. When readers want to catch up on the details of a specific news event, most of them use a search engine to collect news and understand the whole story. It usually takes readers a considerable amount of time to sort out the causes and effects of the event. The general method of online news provision aggregates and organizes the content of news articles from a large number of events and presents the content to readers. Most of this type of information is manually organized. To solve these problems, this study proposes an automated method of news curation. First, we extract the topics from the event data set and use word sequences to find the sequence of topic transfer through a hidden Markov model. Second, we calculate the strength of the topic and the variation in the strength to detect important time points during the development of the news event. Finally, a concise summary is generated at each time point. This paper combines two characteristics, chronology and summary, to design a curation method that can effectively help readers quickly grasp the context of a news event. The experimental results show that the method has good performance in each module, such as the detection of the important phases of events and the creation of the news summary.

Keywords Content curation · Automatic news summary · Event story · Topic detection · Document summarization · News event

✉ Hei-Chia Wang
hcwang@mail.ncku.edu.tw

Chun-Chieh Chen
scout.chen@gmail.com

Ting-Wei Li
r76064124@mail.ncku.edu.tw

¹ Institute of Information Management, National Cheng Kung University, Tainan, Taiwan

² Center of Innovative Fintech Business Models, National Cheng Kung University, Tainan, Taiwan

³ Department of Information Management, National Development Council, Taipei, Taiwan

1 Introduction

The internet has changed behavior in terms of how information is received. People currently tend to obtain information online, such as through e-books, e-papers, and e-magazines, rather than physical paper resources [10, 23]. The internet has also made online news media more popular. According to the “2017 Digital News Report” published by the Reuters Institute for the Study of Journalism (RISJ) [28] (see Fig. 1), the proportion of Americans reading online news is roughly the same as that watching television news in 2012, and it has continued to grow. In contrast, the proportion of people who read traditionally printed paper news has decreased.

[25] pointed out that online news media has been an important resource that people use to digest information, but readers still face an information overload problem. Currently, there are many news websites, and each website has different page designs, such as sections for news classification and most-read stories, so that readers can easily access and browse various news articles. On the other hand, some news events are not one-off but ongoing events (such as the ongoing coronavirus disease 2019 (COVID-19) event). As time goes on, there will be new developments, and subtheme reports may also emerge. Taking the Thailand cave incident in June 2018 as an example, in addition to the initial news of the incident, related news about support, the rescue process and casualties was also reported continuously. At the same time, as an event evolves, a vast number of readers and netizens will also discuss it enthusiastically, and the topic will continue to be an item of interest.

If readers want to learn about continuous news events such as those mentioned above and they perform an event-based search, most of them still use keyword-based searches. Readers enter keywords and search for news events that they want to follow on search engines or news sites, but the results often lead to the following problems [3, 24, 32]:

1. The number of search results is large. Since general search engines or news websites return all news articles through keyword matching. Some excellent search engines have considered user information, search context, and regional search to improve search accuracy. However, general search engine searches often bring about a series of problems such as inaccurate search, excessive amount of information data, and information that does not meet the search needs of users [24]. Some search engines may only have recent information or not relevant information. They need to spend increasing amounts of time

Fig. 1 Sources of news in the USA from 2012 to 2017 [28]

- reading multiple stories [5]. Therefore, readers must read the articles one by one and perform time-consuming filtering to understand the complete phases of an event.
2. Search results are highly sensitive to keywords that have homonym problems. Because readers have different levels of understanding of news events, the keywords used in inquiries may have homophones or homographs, and often containing noise articles. “The 2018 Thailand Football Team Trapped Incident” as an example. When an event cannot be confirmed in advance, users used keywords “Thailand”, “football team” and “incident” to collect news articles, and often collect articles containing noise (i.e., Thailand football team history). It could not collect all the news about the ins and outs of the event effectively.
 3. The search results may have a high news repetition rate.

Given these problems, people may be overwhelmed by news that is not a one-off story, making it difficult to obtain a theme development overview of a news event. This problem can be solved through content curation. Content curation is a thematic integration of online digital content. It activates digital content through collection, filtering, organization, etc., and presents it to people who seek specific subject knowledge [4, 19]. The application of content curation to the news field aims to collect online news articles on specific news issues or events and to integrate the relevant event development overview to readers.

Content curation can be implemented in five modules: aggregation, distillation, elevation, mashup, and chronology [6]. Among them, chronology is most important for the development of news events. [20] highlighted that the analysis of temporal information is useful in a wide range of information retrieval applications and is often an essential component in text understanding. Additionally, [16] defined the term *breakpoint*, which is a time point at which decisive changes occur in the development of an event. Breakpoints can manifest the key phases of a news event and depict the outline of the whole story. On the other hand, in the field of news article analysis in the past, automatic document summarization has been one of the active research areas. [30] pointed out that summaries can help to address information overload. Document summarization merges several documents discussing the same news event and deletes and filters useless or repetitive information to obtain streamlined and focused content [15]. There are two ways to generate abstracts: abstract summarization and extractive summarization. The abstractive summarization is designed to retain the most essential parts of the original document and to regenerate the summarization after understanding the content of the document. The summarization generated by the extractive summarization method is a simplified version that expresses the important information of the original document, retaining the main content and reducing redundant parts. Consequently, extractive summarization is a better approach than abstract summarization [9]. This research data set is derived from news articles originating from different sources to reduce the repetition rate of the same event news and to concisely present the information of the original document. Therefore, this study used an extractive multidocument summarization method to generate news summarization.

In recent years, there have been research methods related to news events (as shown in Table 1). [21] proposed the EventX algorithm, which uses a semisupervised scheme to extract event development combined with a clustering algorithm to find events, but it takes time to label the data. [14] used a hidden Markov model to find the relationship between news frames and sociopolitical events but did not consider automatically summarizing the content of the event summary. [31] proposed using clustering to find each event stage in a multidocument summarization approach to generate a simple and clear summary, but fewer time series were considered. Therefore, this paper proposes a breakpoint concept combined with an automatic content curation method of news events.

Table 1 Research methods for news events

Author	Year	Model
Hopp et al. [14]	2020	Hidden Markov model
Liu et al. [21]	2020	Clustering model, LDA
Wang et al. [31]	2018	Clustering model, summarization

For the retrieval of specific news events, most news websites return a substantial list of news articles. Some of the news is not a one-day event; it may change over many days. To help audiences interested in the progress of an event, we propose an event curation approach that ultimately produces a concise summary of the article. This study combines the two characteristics of a time series and a summary, and the experimental results show that the curation indicators of events obtain a score of 4.5 or more on a scale where the highest score is 5. In addition, the recall-oriented understudy for gisting evaluation (ROUGE) metrics of the automatic news summary increased to 65%. To achieve the above goals, this study aims to collect all the news events as a time series and use breakpoints to detect the news of the development of different events. Furthermore, an automated summary system is established for the progress of each phase of major events according to the content of the event development. This framework collect all news articles regularly for preprocessing, and summarizing the content of events in advance. When the user queries hot events, an overview of the event development will be extracted or performed from the news events that have been collected.

The features of this method include the following two points:

1. We use the concept of a *breakpoint* to analyze the context of news events in a sequential manner.
2. We automatically summarize each *breakpoint* of the news event.

2 Literature review

2.1 Content curation

Curating was originally used in the art field, where it denotes the behavior of curators selecting items for collection and display in museums or galleries. However, as the internet has brought a very large amount of information, curation has been used for digital content, such as digital curation or media curation. Content curation has been applied in many areas, such as news, e-commerce, community information, and education. In addition, content curation is not limited to text form; any multimedia, such as pictures, audio and video, can be the content or the presented result of the content curation. Unlike traditional journalists, who mainly produce content, content curation in the news field focuses on editing and aims to present more organized content to readers. It filters out duplicate and unnecessary content from large numbers of news articles to help readers summarize what they truly need. A practical application of the content curation of news events is the media website Newsy (see Fig. 2). Newsy collects the news on popular topics or issues from authoritative sources, then analyzes them and adds some different perspectives and finally presents them in a video form.

Content curation can be divided into three types depending on the curator [13]:

Fig. 2 Newsy

1. Social curation. This mainly operates through a large number of users in the community who discuss, collaborate on and organize interesting content. For example, recently, new social platforms were created with different business models, and curators could contribute by adding value to the content produced by other users [11]. However, if the curatorial community is small, it will not be able to provide a good response.
2. Expert curation. There can be a group or a single curator who curates a specific field, and the curator must be an expert in the field.
3. Algorithms. Unlike the above two types of curation, in which human participation is required, the goal of algorithms is to conduct content curation through automated methods without human intervention. This type is the focus of our research.

There are some studies on the content curation of algorithms. Most have focused on event detection or document summary, and few studies have suggested a complete automated curation process. [32] used the conditional random field model (CRF) to identify theme sentences. This method did not consider that a theme may last for a long time. For example, the cave rescue by the Thailand football team occurred on June 23, 2018 and lasted for 18 days. Finding appropriate breakpoints for different themes is needed in curation. To build content curation, our research designs a breakpoint finder and analyzes a large number of news articles to present readers with text-based event curation results. We expect that this can help readers quickly and completely understand the development of a whole news event.

2.2 Topic detection

Topic discovery is the premise of event evolution analysis [22]. Topic detection is one of the subtasks of topic detection and tracking (TDT). TDT tasks originated in 1996 as a program sponsored by the Defense Advanced Research Projects Agency (DARPA) to investigate technologies that find and focus on new events in broadcast news [2]. In TDT, a topic is defined as a specific event or concept used to describe a group of related articles. Usually, it is a collection of keywords that are descriptive and can be agglomerated into a similar theme. The task of detecting the theme involved in a corpus (e.g., news, emails, tweets) is topic detection. To date, many scholars have studied this field. Observing the previous related research, the

techniques applied in topic detection can be roughly divided into two categories: nonprobability models and probability models [12, 34]. A nonprobability model detects the topic by methods such as document clustering and graph analysis. A probability model uses probability theory to simulate the uncertainty in the data. For example, a topic can be modeled as a multinomial distribution of words. A probability model describes a set of possible probability distributions of the observed data, and the goal is to learn the best distribution from it.

Probability models have many applications, including comparative text mining (CTM), contextual text mining (CtxTM), and topic sentiment analysis (TSM). The CTM problem was proposed by [33]. As shown in Fig. 3, the CTM task is to discover a common topic (or background theme) for the entire data set and a number of specific latent topics. Similar to the topic model, each document is a mixed model of the common topic and the latent topics, while the latent topics form a word distribution. CTM can mine topics among multiple data sets, so there are many different applications of it, such as summarizing similar products and comparing opinions on specific topics.

However, this study collects a data set based on news events, and each data set can be divided into multiple subsets according to the time interval. As mentioned above, CTM is suitable for our research situation. The subsets after segmentation belong to the same news event, so it is more effective to detect the remaining potential topics after the common background topic is extracted.

2.3 Document summarization

Automated document summarization can handle the problem of information overload and has been an active research area for many years. It can be divided into single-document summarization and multidocument summarization according to the number of original documents. Single-document summarization focuses on making the content more concise and eliminating useless information. Multidocument summarization aims to select information from a

Fig. 3 The CTM model [29]

collection. In addition to streamlining and filtering unnecessary information, it is necessary to effectively avoid duplicate information that appears in different documents.

Document summarization involves creating a rich summary from a document by distilling and extracting the most relevant parts of the text [1]. Document summarization can be roughly divided into abstractive summarization and extractive summarization. In the field of automatic document summarization, most studies focus on extractive summarization [26]. An abstract generated by the extractive summarization method is a simplified version of the important information that represents the original document set [18]. It retains the main content and reduces the redundant part so that the reader can read and understand more quickly. The method of extractive summarization often involves the following steps: (1) intermediate representation, (2) scoring sentences, and (3) selecting summary sentences. First, the document needs to be represented in a structured form. The second step estimates which sentences are most relevant based on the representations created and gives each sentence a score as a measure of its relevance. Finally, the highest-ranked sentences are chosen as the representative sentences to produce a summary [27].

Although the research on document summarization for news has developed widely, it is not applicable to this study. This study is intended to produce a summary of the various news documents that have been cut according to time intervals and include multiple news articles from different sources. Therefore, we still adopt the traditional multidocument summarization method to reduce the repetition rate of news articles that explain the same event and to succinctly present the highlights that occur in the time interval. In general, the performance of extractive summarization is still far superior to that of abstractive summarization [9, 17]. For these reasons, this study also uses the main steps of extractive summarization to design this content curation system.

3 Research methods

In this section, this research proposes an automated content curation method for news events based on the concept of breakpoints. It can divide the news event into several phases and automatically summarize the content of the event evolution.

3.1 Framework

This framework includes a data collection and preprocessing module, filtering module, breakpoint detection module, and curation summarization module. The data collection and preprocessing module are used to collect all news articles and perform word segmentation and stopword deletion. A filter processing module is then used to remove the noise articles from the collected data set. The breakpoint detection module is used to find important time points in the development of events. Finally, the curation summarization module is used for multidocument summarization between each breakpoint. The framework is shown in Fig. 4. Each module is described in detail in the following sections.

3.2 Data collection and preprocessing

Since many news websites provide hot topics on the first page, this study automatically collects hot topics and tracks these events. After collecting the popular news events, the program simulates the process of readers searching for the events on the news website. This

Fig. 4 Method framework

module crawls the news of multiple news websites through the keyword combinations of the event and compiles the news articles into an event data set.

For each news article, the news headline ($title_i$), published time ($date_i$), content ($content_i$) and news keywords ($keyword_i$) are collected. The publication time is essential since the time information can be used to identify whether the event is one-off news. In addition, the article keywords represent the core theme of the news, so we also collect the keywords that the news website offers.

Then, the unstructured content of the news article undergoes preprocessing, including stopword deletion and word segmentation. The word segmentation step converts the news headline and the content into a word collection.

3.3 Filter processing

Since our data collection module simulates the way readers search for events on news websites, two problems were found in the collected news event data set:

1. The search engines of news websites currently mostly use keyword matching and return a large article list. The articles in the list do not all relate to the material of the news event; some of them are noise articles that are not related to the event.
2. A keyword search for an article may return nonevent-related articles. Taking the Thailand cave incident as an example, if the period of the event cannot be confirmed in advance and a keyword matching search is used, noise articles are often collected. Figure 5 shows an example of a noise article. To solve this problem, we use a filter processing module to remove the noise articles from the collected data set and enable the subsequent analysis to achieve a better result.

Fig. 5 An example of a noise article on the event of the Thai soccer team cave rescue in 2018, retrieved in 2019

The filter processing module first extracts the article keywords of each news item. It calculates the occurrence frequency of all article keywords in the data set and obtains the event keyword set (*Event_kword*), which is composed of the keywords with the top *N* frequencies. Then, it compares the keywords of each article with the event keyword set and sets the threshold α . If $(|keyword_i \cap Event_kword|)/|keyword_i|$ is lower than α , then article *i* can be regarded as a noise article and removed from the data set. After filtering by the module, the original news event data set retains the important news material that is truly related to the event. The new data set is defined as $D = \{d_1, d_2, \dots, d_i, \dots\}$, where d_i is the filtered *i*-th news item.

3.4 Breakpoint detection

After preprocessing and filter processing, event-related news set *D* is collected. Then, the development period *T* of the event is determined from the earliest published time to the latest published time in the data set. [16] mentioned that the theme of a news article exhibits a significant change between a *breakpoint* and the time point before it. Thus, this study further cuts the time period of an event into several intervals, represented as $T = \{t_1, t_2, \dots, t_n, \dots\}$, where t_n is the *n*th time interval, and the time interval of this study was analyzed with one day as a basic unit. Likewise, the news event data set is also divided into subsets based on the time interval $\{D_{t_1}, D_{t_2}, \dots, D_{t_n}, \dots\} \subseteq D$. The subset D_{t_n} means the articles collected in the time interval t_n , and *D* is the input of the *breakpoint* detection module. The *breakpoint* detection module includes four steps: topic extraction, construction of the hidden Markov model, determination of the theme strength and determination of the theme variation.

First, background themes z_B and K-specific latent topics $z_1, z_2, \dots, z_k, \dots, z_K$ from the news event data set are obtained by CTM [33]. A topic in the topic model is composed of a set of word distributions, and a background topic is defined as having high-frequency but low-information words. Compared to a background topic, z_k is a more meaningful and distinctive topic. In the CTM model, each word in the news article is represented as a mixture model of the background topic and each latent topic. The word w in the news article d_i is represented as follows:

$$p_{d_i}(w) = \lambda_B p(w|z_B) + (1 - \lambda_B) \sum_{k=1}^K \left[p(z_k|d_i) p(w|z_k) \right]$$

where λ_B represents the mixture weight of the background topic z_B , $p(w|z_B)$ is the probability that word w belongs to background topic z_B , and the conditional probability can be estimated according to the frequency with which w appears in the data set and the frequency of the other words. The following formula is used:

$$p(w|z_B) = \frac{\sum_{d_i \in D} c(w, d_i)}{\sum_{w' \in V} \sum_{d_i \in D} c(w', d_i)}$$

In this formula, V denotes the set of all words in D , and $c(w, d_i)$ represents the number of words w occurring in article d_i . $p(z_k|d_i)$ is also a mixture weight, which indicates the probability that d_i belongs to topic z_k and $\sum_{k=1}^K p(z_k|d_i) = 1$. We represent the topic and the mixture weight in the topic model as $\Lambda = \{z_k, p(z_k|d_i)\}$ and use the log-likelihood to estimate the parameter:

$$\log p(D|\Lambda) = \sum_{d_i \in D} \sum_{w \in V} \left[c(w, d_i) \log \left(\lambda_B p(w|z_B) + (1 - \lambda_B) \sum_{k=1}^K (p(z_k|d_i) p(w|z_k)) \right) \right]$$

In the parameter learning part of the topic model, this work uses the expectation maximization (EM) algorithm for estimation. As the following formula shows, the parameter updates and iterates until convergence. Finally, it yields two probabilities, $p(z_k|d_i)$ and $p(w|z_k)$.

$$\begin{aligned} p(z_{d_i, w} = z_k) &= \frac{p^{(n)}(z_k|d_i) p^{(n)}(w|z_k)}{\sum_{k'=1}^K p^{(n)}(z_{k'}|d_i) p^{(n)}(w|z_{k'})} \\ p(z_{d_i, w} = z_B) &= \frac{\lambda_B p(w|z_B)}{\lambda_B p(w|z_B) + (1 - \lambda_B) \sum_{k=1}^K p^{(n)}(z_k|d_i) p^{(n)}(w|z_k)} \\ p^{(n+1)}(z_k|d_i) &= \frac{\sum_{w \in V} c(w, d_i) (1 - p(z_{d_i, w} = z_B)) p(z_{d_i, w} = z_k)}{\sum_{k'=1}^K \sum_{w \in V} c(w, d_i) (1 - p(z_{d_i, w} = z_B)) p(z_{d_i, w} = z_{k'})} \\ p^{(n+1)}(w|z_k) &= \frac{\sum_{d_i \in D} c(w, d_i) (1 - p(z_{d_i, w} = z_B)) p(z_{d_i, w} = z_k)}{\sum_{w' \in V} \sum_{d_i \in D} c(w', d_i) (1 - p(z_{d_i, w'} = z_B)) p(z_{d_i, w'} = z_k)} \end{aligned}$$

After topic extraction, the hidden Markov model (HMM) is used to determine the transformation sequence of the topic in the news article. Each news article d_i is treated as a word stream, and the word streams of all articles in D_{in} are concatenated as a word sequence. The word sequence will be used as the observation sequence of the HMM, and a topic extracted in the previous stage is regarded as a hidden state. We create an HMM, as shown in Fig. 6. Each word in the word sequence corresponds to a topic, and the emission probability distribution is $p(w|z_k)$. The initial state probabilities and transition

Fig. 6 Topic shift model using an HMM

probabilities are estimated by the Baum-Welch algorithm [29]. The Viterbi algorithm [29] is used to solve the decoding problem and predict the transformation of the topic sequence.

This study then calculates the topic transformation sequence and aggregates it into a value defined as the theme strength. The theme strength of topic z_k in time interval t_n is represented by $\sigma_{t_n}(z_k)$. The theme strength is calculated as follows: The numerator is the number of words in the sequence that belong to topic z_k , and the denominator is the length of the word sequence of D_{t_n} . Finally, each time interval t_n will produce a distribution of the theme strength $\{\sigma_{t_n}(z_1), \sigma_{t_n}(z_2), \dots, \sigma_{t_n}(z_K), \sigma_{t_n}(z_B)\}$. The greater the value of the theme strength, the more likely it is that this time interval belongs to the topic.

$$\sigma_{t_n}(z_k) = \frac{|\{w | w \in D_{t_n} \cap w \in z_k\}|}{|D_{t_n}|}$$

Since a *breakpoint* refers to the occurrence of a decisive change during the development of the event, the topics will change in the time interval before and after the *breakpoint*; for example, there is a breakpoint between when the cave rescue search starts and when the boys are found. In view of this, this study detects the time interval in which the *breakpoint* is located by calculating the variation of the theme strength. According to the results of the previous stage, each time interval has a theme strength distribution. We use the Jensen-Shannon divergence (JS divergence) [8] to measure the difference between the time intervals. The theme variation (TV) between t_n and t_{n+1} is defined as $TV_{(t_n, t_{n+1})}$, and it is calculated as follows:

$$TV_{(t_n, t_{n+1})} = \frac{1}{2} \sum_{k=1}^{K+1} \sigma_{t_n}(z_k) \log \frac{\sigma_{t_n}(z_k)}{m_{z_k}} + \frac{1}{2} \sum_{k=1}^{K+1} \sigma_{t_{n+1}}(z_k) \log \frac{\sigma_{t_{n+1}}(z_k)}{m_{z_k}}$$

In the formula, $m_{z_k} = \frac{1}{2} [\sigma_{t_n}(z_k) + \sigma_{t_{n+1}}(z_k)]$, and the number of topics is $K + 1$, including one background topic z_B . After calculating the TV of each adjacent time interval, it is then determined whether the time interval is a *breakpoint* by the following condition.

$$\begin{cases} TV_{(t_{n-1}, t_n)} > TV_{(t_{n-2}, t_{n-1})} + \gamma \\ TV_{(t_{n-1}, t_n)} > TV_{(t_n, t_{n+1})} + \gamma \end{cases}$$

When the TV value is small, the difference in the theme between the two time intervals is small, which means that the same topic continues. If the two conditions are met, the time interval t_n is determined as a *breakpoint*. However, the first, second and last time intervals during the event period cannot be detected by the above conditions. In this part, we compare the maximum theme strength to the average of the theme strength in the interval. Following the formula below, if the maximum theme strength value is greater than the average value, the time interval is also regarded as a *breakpoint*.

$$z_k \in Z, i = 1, 2, N \max(\sigma_{t_i}(z_k)) > \frac{\sum_{z_k \in Z, i=1,2,N} \sigma_{t_i}(z_k)}{K+1}$$

3.5 Curation summarization

After detecting all *breakpoints* during the development of the event, the study conducts a multidocument summarization between *breakpoints*. The entire event was aggregated into a concise textual description in the time interval in which progress occurred and finally formed the result of the event curation. The summary module first regroups the news event data set into $\{D_{b_1}, D_{b_2}, \dots, D_{b_m}, \dots\} \subseteq D$ based on the *breakpoints*. D_{b_m} represents all the news articles at *breakpoint* b_m . D_{b_m} is used as the input of the summary module. The first step is to extract the sentences from D_{b_m} , cluster the sentences and finally extract the representative sentences of each group to form the final summary.

Before clustering, the sentences must be represented by a vector matrix. We extract all the sentences from the news data set of the *breakpoint* and store them with the sentence set $S = \{s_k | s_k \in D_{b_m}\}$. Then, the sentences are converted into vectors by one-hot encoding. After the sentence expression is completed, this study uses the spherical K-means proposed by [7]. Spherical K-means is a measure of the similarity between data points based on the cosine similarity. The similarity between two sentences is calculated as follows: The directional characteristics of the vector are considered. When the similarity is greater than 0, $\text{sim}(s_i, s_j) > 0$, which means that the two sentences are similar. Finally, similar sentences are clustered into a group.

$$\text{sim}(s_i, s_j) = \frac{\overline{s_i} \bullet \overline{s_j}}{\|\overline{s_i}\| \times \|\overline{s_j}\|}$$

The last stage of this module is to select the representative sentences. We select them from the sentence cluster obtained in the previous stage and choose the representative sentences of each cluster. First, we establish a topic signature word set (TSW) for the *breakpoint*. The topic signature word selection follows the theme strength of the *breakpoints*. Each *breakpoint* has a theme strength distribution $\{\sigma_{t_n}(z_1), \sigma_{t_n}(z_2), \dots, \sigma_{t_n}(z_K), \sigma_{t_n}(z_B)\}$. From this distribution, we take the top two high-strength themes and extract ten words with the highest $p(w | z_k)$ from each theme into TSW . Next, each sentence in the cluster is stored as a bag-of-words model, and the stopwords are removed to retain only useful words. The bag-of-words model is represented by SW . In this stage, we compare TSW with SW . We use the Jaccard similarity to calculate the similarity between the two, as given in the formula below. We use the similarity as the score to rank the sentences. The higher the similarity is, the higher the ranking. Finally, the representative sentences of each cluster are removed and form a news summary of the *breakpoints*.

$$\text{Jaccard}(\text{TSW}, \text{SW}) = \frac{|\text{TSW} \cap \text{SW}|}{|\text{TSW} \cup \text{SW}|}$$

4 Experiments

4.1 Data set

This study involved the most popular news publishers according to the 2018 Digital News Report (see Fig. 7), which was published by the RISJ, including Ettoday.net, Apple Daily and United Daily of Taiwan. In Fig. 7, Yahoo! News is not selected since it is a news aggregator, not a publisher.

Additionally, five popular news events between 2013 and 2019 are selected for evaluation. These events cover a wide range of issues, including disasters, conflicts, accidents and social movements. For each news event, we collected all the search results from the above three news websites. Table 2 provides a brief description of our data set.

Moreover, we also constructed a gold standard. For each news event, we collected human-written timeline overviews from six authoritative online media, including five news agencies (Ettoday.net, Apple Daily, United Daily, Nownews, and Newtalk) and Wikipedia. Each human-written overview contained a list of time points and a corresponding summary. However, considering that each human-written overview has its own focus on the event, we selected the time points that were present in at least three human-written overviews as the gold standard of the *breakpoints* to ensure consistency. Once a *breakpoint* was selected, all human-written summaries for that date were collected to form the gold standard of the *breakpoint* summary.

Fig. 7 The use of online news media by Taiwanese respondents in 2018

Table 2 A brief description of our data set

News event	Brief introduction	Time span	# of News items
Thailand cave incident	The 12-member “Wild Boars” soccer team and their coach were trapped in the Tham Luang cave because heavy rains flooded the cave.	20 days	203
Sun An-Tso event	A Taiwanese exchange student accused of threatening a shooting at Delaware County High School was to be deported.	8 and a half months	592
Hung Chung-Chiu event	The death of an army conscript in Taiwan caused public concern due to suspected bullying, abuse and other military scandals.	2 and a half years	1820
The election at the National Taiwan University (NTU) president	During the election of the NTU president in 2018, a series of disputes about the presidential selection system and academic autonomy arose.	1 year	562
Pilot strike against China Airlines (CAL)	A pilot strike was launched by the Pilots Union Taoyuan, and it led to a series of negotiations between the union and CAL.	7 months	475

4.2 Metrics

The experiment is divided into two parts. The first part is the effect analysis of the system used to generate *breakpoints* and summaries, and the second part is the manual evaluation of the overall curation result. In the *breakpoint* detection step, this study uses the evaluation indicators precision, recall and F1-measure to verify the effect of the *breakpoint* generated by the system. We compare the system *breakpoint* to the gold standard and classify the numbers into four categories, namely, TP, FP, FN and TN, defined as follows. TP indicates that a *breakpoint* detected by the system also exists in the gold standard. FP indicates that a *breakpoint* detected by the system does not belong to the gold standard. FN represents a time point that exists in the gold standard but is not detected by the system. TN indicates that the system and manual tags are not considered *breakpoints*. These four categories can be used to calculate the accuracy of *breakpoint* detection using the following formula:

$$\begin{aligned}
 \text{Precision} &= \frac{TP}{TP+FP} \\
 \text{Recall} &= \frac{TP}{TP+FN} \\
 F1 - \text{measure} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
 \end{aligned}$$

The precision evaluates how many detected *breakpoints* are truly important developments in the event; the recall represents the number of detected *breakpoints* among all *breakpoints* that exist in the gold standard; and the F1-measure is a comprehensive comparison combining the precision and recall.

For the generated summary evaluation, we use the ROUGE to compare the automatically generated summary with the reference summary, which is human written, and calculate the co-occurrence frequency of n-grams between them. The formula is given below, where gram_n is an n-gram and $\text{Count}_{\text{match}}(\text{gram}_n)$ indicates the number of times that the system-generated summary and the reference summary both contain gram_n . The larger the ROUGE-N is, the closer the two summaries are.

$$ROUGE - N = \frac{\sum_{S \in \{\text{reference summary}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{reference summary}\}} \sum_{gram_n \in S} Count(gram_n)}$$

The second step is a manual evaluation of the curation results. We refer to the indicators proposed by [26]. In this paper, we use six indicators to measure the quality of the curation result: usefulness, coherence, referential clarity, nonredundancy, focus, and overall. The evaluation method is mainly based on the findings of researchers in the related domain of information management. The indicators are defined as follows:

1. Usefulness. This is used to understand whether the curation result truly gives readers information about the development of the event.
2. Coherence. This is used to check whether the structure of the curation result is reasonable and whether it seems fluent to the readers.
3. Referential clarity. This tracks whether the reader can clearly know who or what appears in the curation result.
4. Nonredundancy. Unnecessary duplicate words or sentences should not occur in the curation result.
5. Focus. This tracks whether the curation result includes sentences or information that are not related to the development of the event.
6. Overall. Finally, the quality of the curation result is evaluated from the perspective of the reader.

4.3 Parameter settings

The methods proposed in this study need to determine the parameters below in advance. Therefore, we aim to find the best system parameters for the subsequent experiment by the methods below.

4.3.1 Parameter 1. The threshold α of the filter processing module

In the filter processing module, we compare the event keyword set to the article keywords to remove irrelevant articles. $N = 10$ is used in this module; that is, the top ten keywords are added to the event keyword set. After the event keyword set is established, the cut ratio is adjusted according to the threshold α . The lower α is, the more irrelevant articles are retained. In contrast, if α is too high, there may be excessive deletion, resulting in too few documents remaining in the data set. Therefore, we tested the effect of α on the accuracy of *breakpoint* detection.

According to the results, the best F1-measure score was obtained for each event when $\alpha = 0.7$, so α was set to 0.7 in subsequent experiments.

4.3.2 Parameter 2. The mixture weight λ_B of the background topic in the *breakpoint* detection module during topic extraction

In the *breakpoint* detection module, we first conduct topic extraction. This study uses the topic model of the CTM task. It represents each word as a mixture model of the background topic and the latent topics. The mixture weight λ_B controls the influence of the background theme. In

the original CTM model, the data set can contain stopwords. In this case, λ_B should be set to a high value, and the model will automatically eliminate the less informative words from the topic cluster. Conversely, if the data set is concise and most words are informative, then λ_B should be set to a small value.

However, the input data set in this study is a news article collection for the same event. The stopwords are removed during preprocessing, and we desire to use the topic model to display the words of a common theme in the data set. Thus, a lower value will be set. According to the results, when $\lambda_B = 0.2$, the best accuracy is obtained for each event, so the background weight of the topic model in the study is set to 0.2.

4.3.3 Parameter 3. The latent topic number K in the *breakpoint* detection module during topic extraction

As mentioned above, in the *breakpoint* detection module, a background topic and K latent topics are extracted, so it is still necessary to determine the number of latent topics. In the past, the literature that used the same CTM model for the news data set mostly set K values of 5 to 7 for analysis. Therefore, this step is based on a value of 5 to 7 plus or minus 3 to obtain a range of 2 to 10 to find the most appropriate number of topics. According to the results, each event has its own best number of topics, and they do not have the same K value. Therefore, we also calculate the average of the five events. The results show that when $K = 4$, the highest average is obtained; that is, the model extracting four latent topics is relatively good. Therefore, we set $K = 4$ for analysis in subsequent experiments.

4.4 Experimental results

The filter processing module, which is designed to solve the problems caused by the traditional keyword search method, is evaluated first. The experimental results are shown in Fig. 8. Compared with the results of the *breakpoints* detected without the module, *breakpoints* that have been filtered are improved in all of the indicators, precision, recall and F-measure. However, in the “Hung Chung-Chiu event”, the recall decreased. We examined the results and found that in terms of the TP value, the result without filtration has one more *breakpoint* than the result with filtration, but too many *breakpoints* are detected. Some *breakpoints* have a large topic variation but are not truly *breakpoints* for the event. Thus, we use the F-measure, which is the average of the two, and it shows that the filter processing module can truly help improve the quality of breakpoint detection. Moreover, this study further conducted a t test of both unfiltered and filtered samples. According to the experimental results, the p value of the F-measure is 0.0020, which indicates that there is a significant difference between the two (p value < .005). Consequently, it is confirmed that the filter processing module proposed in this study is effective.

Second, we verify the detection effect of the *breakpoint* detection module. This study calculates the strength and variation of the theme and detects the time points that meet the criteria as *breakpoints*. This experiment compares our method with the method that detects *breakpoints* simply by the number of news articles. This quantity method intuitively reflects that when an event has an important development, news agencies will release news items in large numbers. We calculate the number of news articles at each time point in the event data set and set a threshold that indicates the ratio of the number of articles. The time points that occur just before the ratio of the threshold is met are selected as *breakpoints*. In the quantity method, the number of detected *breakpoints* increases with the threshold, but the accuracy may not

Fig. 8 Impact of the filter processing module on breakpoint detection for different events

increase. Therefore, we refer to [16] and set the threshold to 15%; that is, we select a time point with the first 15% of the articles occurring before it as a *breakpoint*.

The experimental results are shown in Fig. 9. The quantity method has a relatively high precision value for all five events, and even in the Thailand cave incident and Sun An-Tso event, the precision value is equal to 1. The main reason for this is that, under the same threshold ratio, the number of time points included in the event also affects the number of *breakpoints* detected by the quantity method. Therefore, for events with short time ranges, it may be that only one or two *breakpoints* are detected and a fairly high precision is obtained. However, a good curation result is expected to be able to detect all *breakpoints* during event development. Therefore, the second experiment also uses the F-measure as the main measure. According to the experimental results, the *breakpoint* detection method of this study is superior to the quantity method for each event.

Third, we verify the performance of the summary module. This study obtains an automatic document summary for each *breakpoint* detected by the system. This experiment compares the system-generated summary with the reference summary of the gold standard and measures the difference by the statistical evaluation indicator ROUGE-N. In our summary module, the number of sentences in the generated summary is affected by the number of clusters. To determine the number of clusters, we count the number of sentences in the reference summary

Fig. 9 Comparison of *breakpoint* detection methods

of each event and take the average number of sentences, five, as the number of clusters. In addition, this study evaluates $N = 1$ and $N = 2$ as the n -gram length in ROUGE- N to calculate the recall rate of the summary.

In the third experiment, the summary method of this study is compared with the method using K-means with the Euclidean distance. The experimental results are shown in Fig. 10. Since the summary generated by this study is based on the output of the above *breakpoint* detection module, the ROUGE- N score is calculated only for the *breakpoints* that were detected correctly. The experimental results present the average ROUGE- N score for each event. According to the results, the summary of our study has better performance in both ROUGE-1 and ROUGE-2 than the compared method.

As mentioned above, we measured the accuracy of the detected breakpoints and the ROUGE scores of the generated summaries but did not measure the practicality of the curation results for humans. We therefore manually evaluate the overall quality of the curation results generated by the system. We use six indicators—usefulness, coherence, referential clarity, nonredundancy, focus, and overall—to score the curation results of each event. Every reader gives each indicator a score for all events after reading the generated curation results. The mechanism of giving scores is a five-level Likert scale, ranging from strongly disagree (1) to strongly agree (5).

The experimental results are shown in Fig. 11. The curation results of each event achieved a high score, 4.5 or more on average, for usefulness, referential clarity, nonredundancy, and focus. The scores for the coherence indicator are slightly lower. The reason is mainly that the summary method of our study only selects representative sentences from each cluster to compose the summary and does not address the order of sentences. Therefore, the reader

Fig. 10 Comparison of automatic news summary methods

may give a low score to the coherence. Overall, the indicators for curation results range from very poor quality (1) to very good quality (5). The results show that the quality of the curation of the five events, from the reader's viewpoint, is good. While search engines have mainly been used by people to understand “what's new” or “what's going on” [5], they cannot summarize the content of each development phase of an event in a sequential manner for major events. Therefore, this study used random interviews to determine whether needs were met by comparing the curatorial mechanism and search engines. The statistics showed that 80% of the interviewed people were interested in using the curatorial mechanism and could quickly understand the full picture of an event's development, whereas the remaining 20% were able to find the latest information about events they had followed through search engines. In other words, most people believe that curation mechanisms are better than search engines for finding the news event information they want. This is especially true when time constraints prevent searching for many news events. Therefore, this research proposes an advantageous curatorial mechanism that automatically summarizes the development phases of events. It is more attractive to readers and helps readers understand complex events [21, 31]. Furthermore, the results of this study show the evolution of events at each time point, which is exemplified by an event summary of the “Thailand cave rescue” that was generated by the proposed method, as shown in Fig. 12.

Fig. 11 Results of the manual evaluation of each curation event

5 Conclusion and future work

With the advancement of network technology and the popularity of mobile devices, the news media industry has gradually transformed to digitalization, and the reading habits of readers have also changed to favor online news. Currently, although news websites offer abundant online news for readers to browse, readers who want to understand the context of specific events still face information overload. The technique of content curation therefore applies to news events. However, most current news event curation methods rely on manual writing, which takes considerable time and effort. In view of this, this study proposes an automated content curation method that combines time information and summary to help readers learn about event development easily.

The proposed method of automated content curation is applied to news events. We expect to help news readers who seek specific event information quickly obtain useful information. In the process of our automatic curation method, we first collect a news event data set by

Fig. 12 Curated results of news events taking the “Thailand Cave Rescue Incident” as an example

simulating readers' searches. Since the search engines of current news websites mainly use keyword matching, if a keyword appears in the content or title, the article will be returned to the reader. This often leads to the website providing the reader with too many articles that are nonrelevant or duplicated, and it is also quite inconvenient for the reader to view the search results. Therefore, a filtering processing module was designed to prescreen the collected news data set before analysis. In the experiment, the effect of the filtering processing module was also verified. We compared the performance of breakpoint detection with and without the module and checked the results with the T test. The results show that adding the filtering processing module can significantly improve the results of breakpoint detection.

In the curation section, this study combines the time sequence and summary to design the presentation of the curated events and automates these elements separately. In the time sequence step, the topic model and an HMM were used to express each time interval according to the strength of the theme and set conditions to find the time intervals with large TV values. In the automated document summary step, this study uses the extractive summarization method to design our summary module. In the stage of selecting sentences from the module, the theme strength and topic word distribution from the breakpoint detection module are used to compare and score the sentences. Finally, an overview of the event is presented to readers.

This study predicts the breakpoints from the topic transfer sequence in the breakpoint detection module. The theme strength and TV at each time point are calculated, and the time point at which the variation is greater than previously is determined to be the breakpoint. In the experiment, the quantity method, which detects breakpoints based on the number of news articles, was compared, and the results showed that our approach was better than the existing method. Since the breakpoint in the gold standard is taken from the artificial edit, the time is usually more precise. Our automatic detection method relies on news articles that are affected by the release time. Therefore, there may be a time delay, and even the time delay of international news may be long. These situations may result in an overall score that is not as good as expected.

The summary module in this study is designed using the extractive summarization method. We adopt the spherical K-means method in the sentence clustering stage and compare it with the K-means method using the Euclidean distance. The results show that the summary method of this study performs well in both ROUGE-1 and ROUGE-2. We also explore the reasons for this, which may be due to the quality of Chinese word segmentation.

This study also measures the curation results from the reader's point of view. The results of the manual evaluation show that the curation results of this study perform well in the six indicators of usefulness, coherence, referential clarity, nonredundancy, focus and overall. The data set used in this study has several limitations: 1) the publication time is essential, 2) news keywords must be available, and 3) news development needs to last more than one day. Therefore, this research method is not applicable to some data sets that do not contain time information and news keywords. At present, most news articles on most websites contain keywords (e.g., BBC and The Guardian). For the data without the keywords, it could refer to Google trends to get keywords to collect news articles. That could be evaluated in the future works of this research. This study proposes suggestions for future research directions regarding automatic content curation for news events. First, the automatic *breakpoint* detection method in this study relies on the release time of news articles. However, the release time is affected by news organizations, and even different news sources may have different release times. Therefore, the time delay may be several days after the actual event development time. To solve this problem, this study proposes two improvement ideas. Before using the condition to detect the *breakpoint*, we can further analyze the adjacent time intervals and conduct merging

or filtering. The second idea is that a news article may contain the time information of the *breakpoint*. If we can extract the time information from the text and then perform *breakpoint* detection from these time points, it may make the detection results more accurate.

In addition, the automatic summarization method of this study uses extractive summarization. We selected the best sentence from each group to compose the summary but did not consider the order of the sentences. In the future, the order of the sentences can be processed. Additionally, future work can aim to generate the summary by the abstractive summarization method so that the summary will be smoother and will provide a better curation result for readers.

Code availability Not applicable.

Funding Ministry of Science and Technology, Taiwan.

Data availability Not applicable.

Declarations

Conflict of interest To the best of our knowledge, the named authors have no conflict of interest, financial or otherwise.

References

1. Alami Merrouni Z, Frikh B, Ouhbi B (2020) Automatic keyphrase extraction: a survey and trends. *J Intell Inf Syst* 54:391–424
2. Allan J, Carbonell J, Doddington G, Yamron J, Yang Y (1998) Topic detection and tracking pilot study: final report. In: *Proceedings of the DARPA broadcast news transcription and understanding workshop*, 1998, pp 194–218
3. Antoniou G, Harmelen FV (2008) *A semantic web primer* (2nd ed.). The MIT Press, Cambridge, MA
4. Burnette-Lemon J (2012) The collector: Pearlrees' Oliver Starr explains how content curation works for both individual users and companies. *Communication World* 29:24–27
5. Chang HT, Liu SW, Mishra N (2015) A tracking and summarization system for online Chinese news topics. *Aslib J Inf Manag* 67(6):687–699. <https://doi.org/10.1108/AJIM-10-2014-0147>
6. Dale S (2014) Content curation: the future of relevance. *Bus Inf Rev* 31(4):199–205
7. Dhillon IS, Modha DS (2001) Concept decompositions for large sparse text data using clustering. *Mach Learn* 42(1):143–175
8. Endres DM, Schindelin JE (2003) A new metric for probability distributions. *IEEE Trans Inf Theory* 49(7):1858–1860
9. Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. *J Artif Intell Res* 22:457–479
10. Giomelakis D, Karypidou C, Veglis A (2019) SEO inside newsrooms: reports from the field. *Future Internet* 11(12):261
11. Guidi B, Michienzi A, Ricci L (2021) A graph-based socioeconomic analysis of Steemit. *IEEE Transactions on Computational Social Systems* 8(2):365–376. <https://doi.org/10.1109/TCSS.2020.3042745>
12. Haribhakta, Y., Malgaonkar, A., & Kulkarni, P. (2012). Unsupervised topic detection model and its application in text categorization. *Proceedings of the CUBE international information technology conference*, 314–319.
13. Herther NK (2012) Content curation: quality judgment and the future of media and web search. *Searcher* 20:30–41
14. Hopp FR, Fisher JT, Weber R (2020) Dynamic transactions between news frames and sociopolitical events: an integrative, hidden Markov model approach. *Journal of Communication* 70(3):335–355. <https://doi.org/10.1093/joc/jqaa015>

15. Hou SL, Lu RQ (2020) Knowledge-guided unsupervised rhetorical parsing for text summarization. *Information Systems*:94. <https://doi.org/10.1016/j.is.2020.101615>
16. Hu P, Huang M, Xu P, Li W, Usadi AK, Zhu X (2011) Generating breakpoint-based timeline overview for news topic retrospection. In: 2011 IEEE 11th international conference on data mining, pp 260–269
17. Huang T-C, Hsieh C-H, Wang H-C (2018) Automatic meeting summarization and topic detection system. *Data Technologies and Applications* 52(3):351–365
18. Huang YX, Yu ZT, Guo JJ, Yu ZQ, Xian YT (2020) Legal public opinion news abstractive summarization by incorporating topic information. *Int J Mach Learn Cybern* 11(9):2039–2050
19. Kanke T (2019) Knowledge curation work in Wikidata WikiProject discussions. *Library Hi Tech* 39:64–79. <https://doi.org/10.1108/lht-04-2019-0087>
20. Li HR, Zhu JN, Ma C, Zhang JJ, Zong CQ (2019) Read, watch, listen, and summarize: multi-modal summarization for asynchronous text, image, audio and video. *IEEE Trans Knowl Data Eng* 31(5):996–1009
21. Liu B, Han FX, Niu D, Kong L, Lai K, Xu Y (2020) Story Forest: extracting events and telling stories from breaking news. *ACM Trans. Knowl. Discov. Data* 14(3) Article 31. <https://doi.org/10.1145/3377939>
22. Liu T, Xue F, Sun J, Sun X (2020) A survey of event analysis and mining from social multimedia. *Multimed Tools Appl* 79(45):33431–33448. <https://doi.org/10.1007/s11042-019-7567-7>
23. Loan FA (2011) Impact of internet on Reading habits of the net generation college students. *International Journal of Digital Library Services* 1(2):43–48
24. Ma X (2020) A new aggregated search method. *Journal of Intelligent & Fuzzy Systems* 38:55–63. <https://doi.org/10.3233/JIFS-179380>
25. Makhortykh M, de Vreese C, Helberger N, Harambam J, Bountouridis D (2020) We are what we click: understanding time and content-based habits of online news readers. *New Media Soc* 23:2773–2800. <https://doi.org/10.1177/1461444820933221>
26. Marujo L, Ling W, Ribeiro R, Gershman A, Carbonell J, de Matos D, Neto JP (2016) Exploring events and distributed representations of text in multi-document summarization. *Knowl-Based Syst* 94:33–42
27. Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Aggarwal CC, Zhai C (eds) *Mining text data*. Springer Science & Business Media, New York, NY, pp 43–76
28. Newman N, Fletcher R, Kalogeropoulos A, Levy DA, Nielsen RK (2017) Reuters Institute digital news report:2017
29. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77(2):257–286
30. Sankarasubramaniam Y, Ramanathan K, Ghosh S (2014) Text summarization using Wikipedia. *Inf Process Manag* 50(3):443–461
31. Wang C, He X, Zhou A (2018) Event phase oriented news summarization. *World Wide Web* 21(4):1069–1092. <https://doi.org/10.1007/s11280-017-0501-x>
32. Xu J, Yang X (2015) Generating the theme overview based on clue chain from online news. In: *Proceedings of the 2015 IEEE international conference on systems, man, and cybernetics*, pp 2730–2735
33. Zhai C, Velivelli A, Yu B (2004) A cross-collection mixture model for comparative text mining. In: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pp 743–748
34. Zhao T, Luo X, Qin W, Huang S, Xie S (2018) Topic detection model in a single-domain Corpus inspired by the human memory cognitive process. *Concurrency and Computation: Practice and Experience* 30(19): 1532–0626